High Speed Computing, Salishan 2002: Next Generation Scalable Network Storage Architecture

Garth Gibson
CTO, Panasas Inc
Assoc Professor, CS & CE, Carnegie Mellon
ggibson@panasas.com, garth@cs.cmu.edu

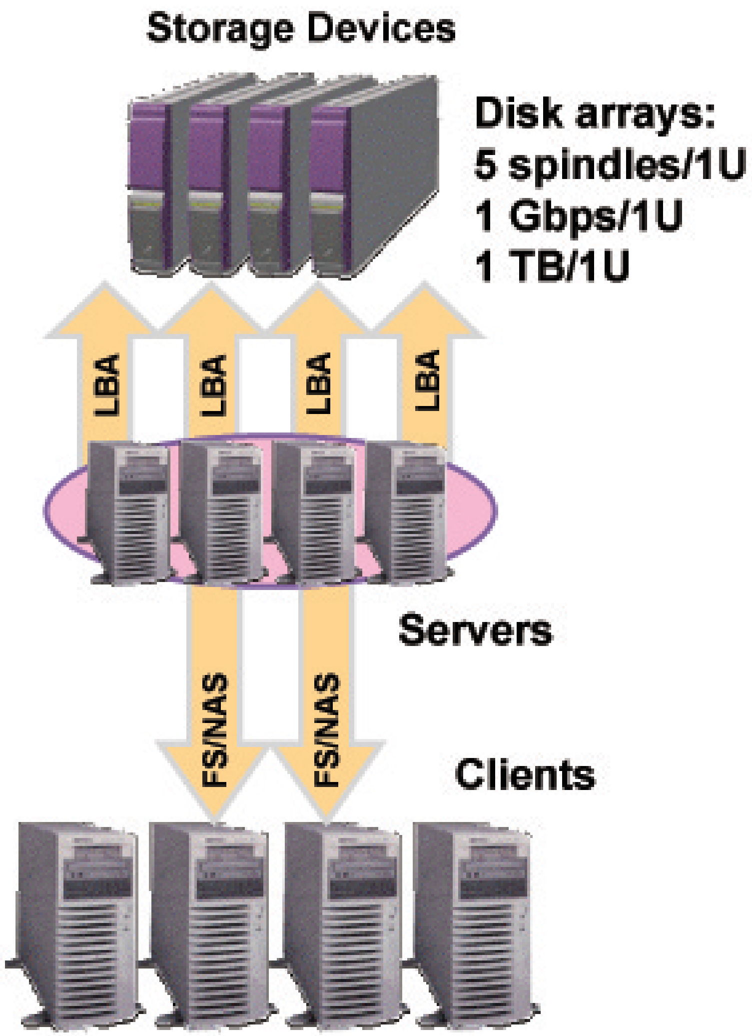
panasas

Technical Market Frustration

- **Needs run years ahead of market influence**
- **Files, systems and parallelism much larger than most**
  - ***Expensive bandwidth requirements!***
  - Rare levels of data density requirements, incremental growth rates
  - Bin packing giant files often falls to users
- **Unique requirements for sharing, reliability and security**
  - End users write programs for shared data -- representation standards
  - Raw numbers of components outpace state of the art FT
  - National security intermingled with international collaboration

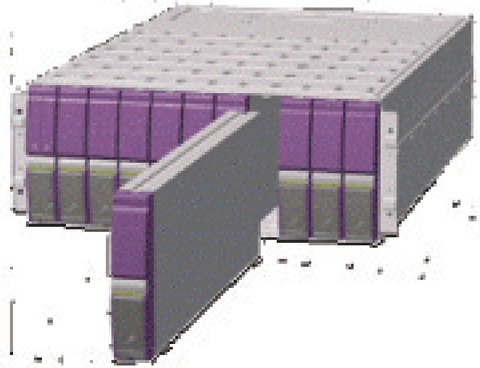
Cluster FS (In-band) Architecture

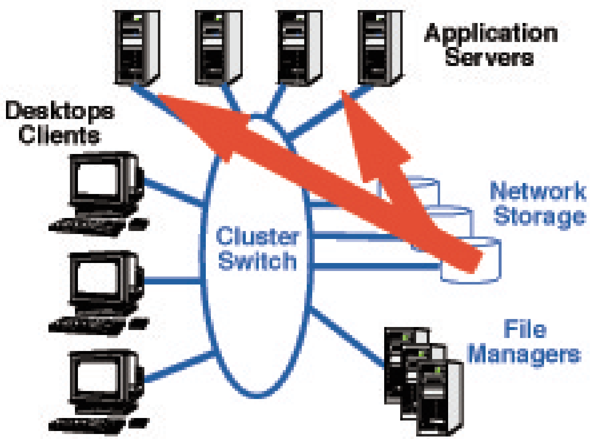
- **Moving data through central file servers limits cluster FS**
  - PCs barely better than disks at moving data, but more expensive
  - Storage vendors amortize server costs with lots of disks
  - Achievable bandwidth very short of raw disk bandwidth (1-4 Gbps/rack)
- **Cluster FS too often derived from single process software**
  - Excessive locking and inter-server data motion
  - Specialized hardware not unusual

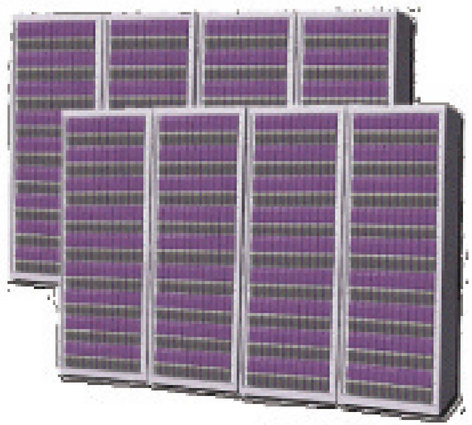


Scale Architecturally

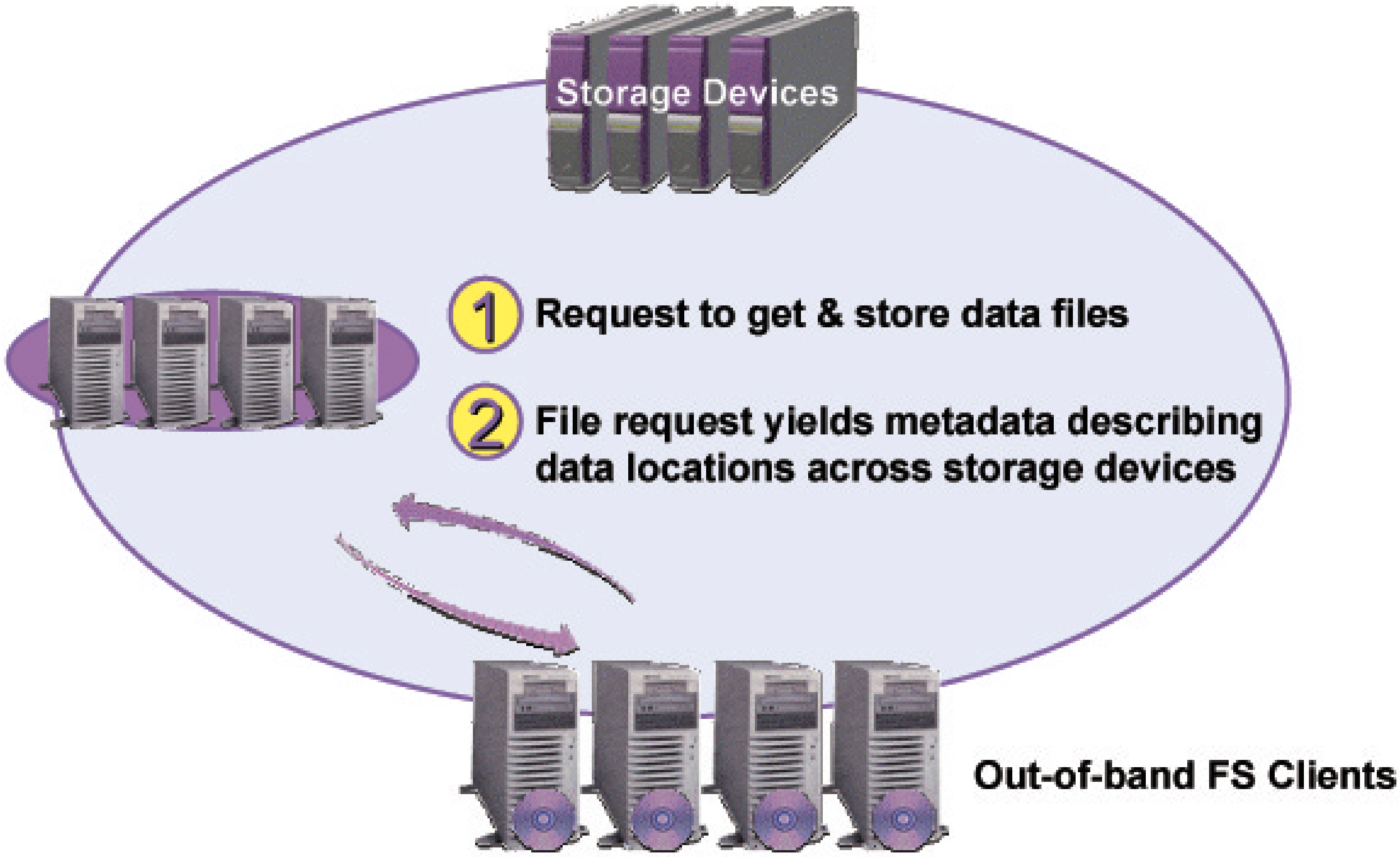
- **Direct, parallel storage access**
- **Commodity technology with integrated functionality**
- **Shared-nothing clusters of data & metadata**





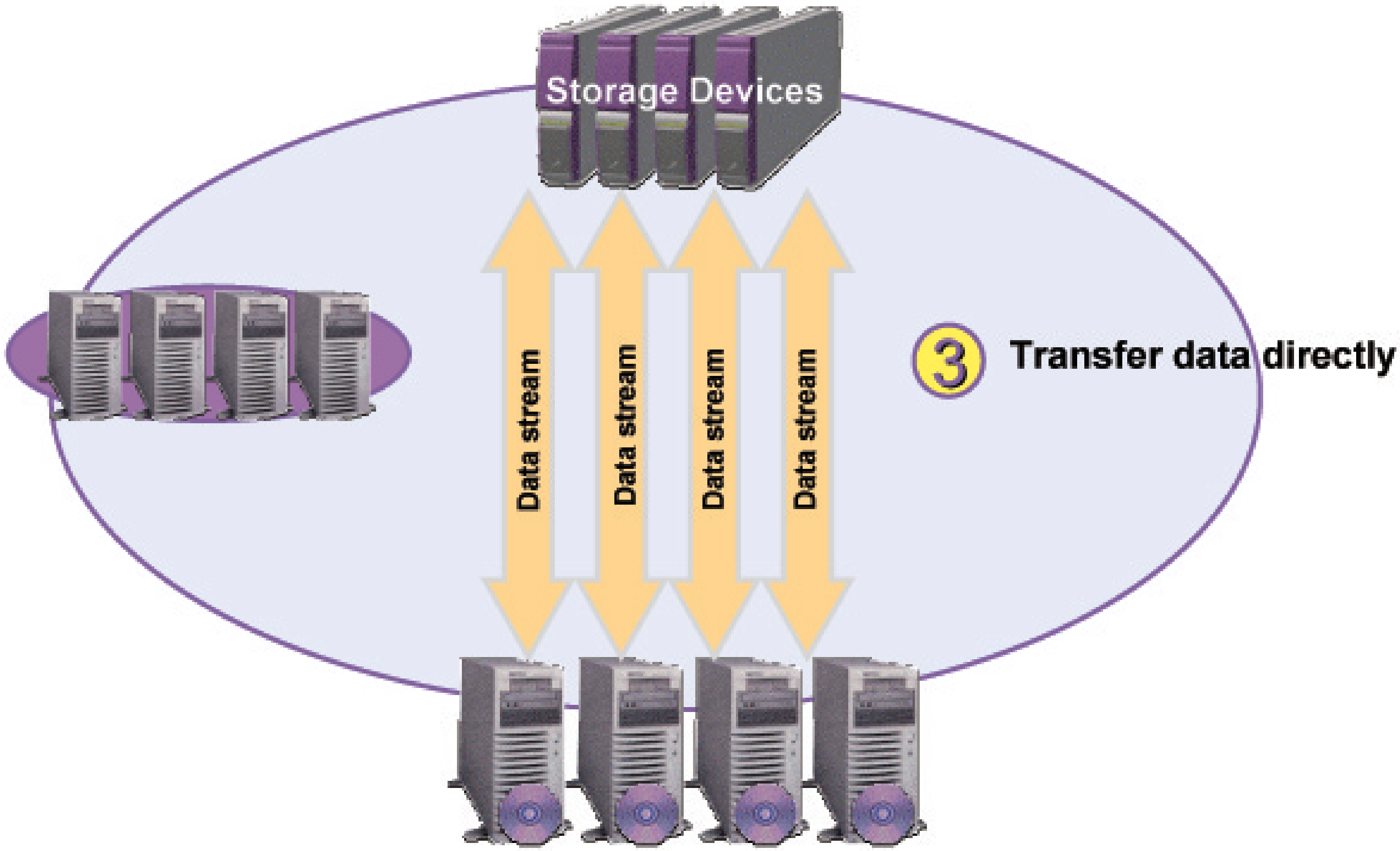


Direct Access for Bandwidth



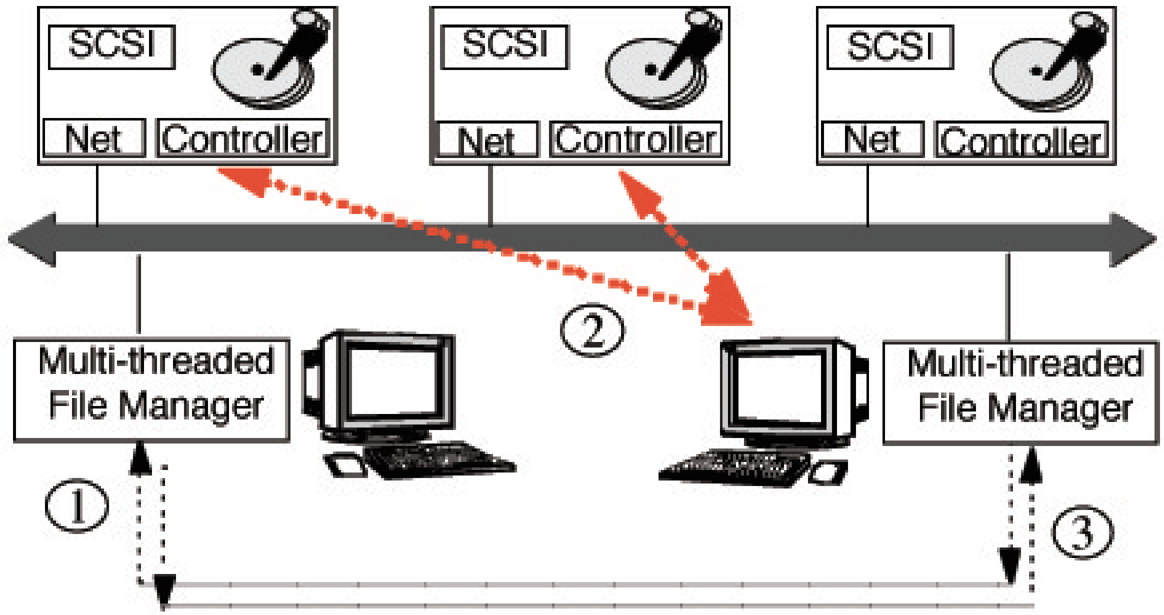
Storage Network = Client Network

Orders of magnitude increase in cost-effective bandwidth



Out-of-band 1: SMP

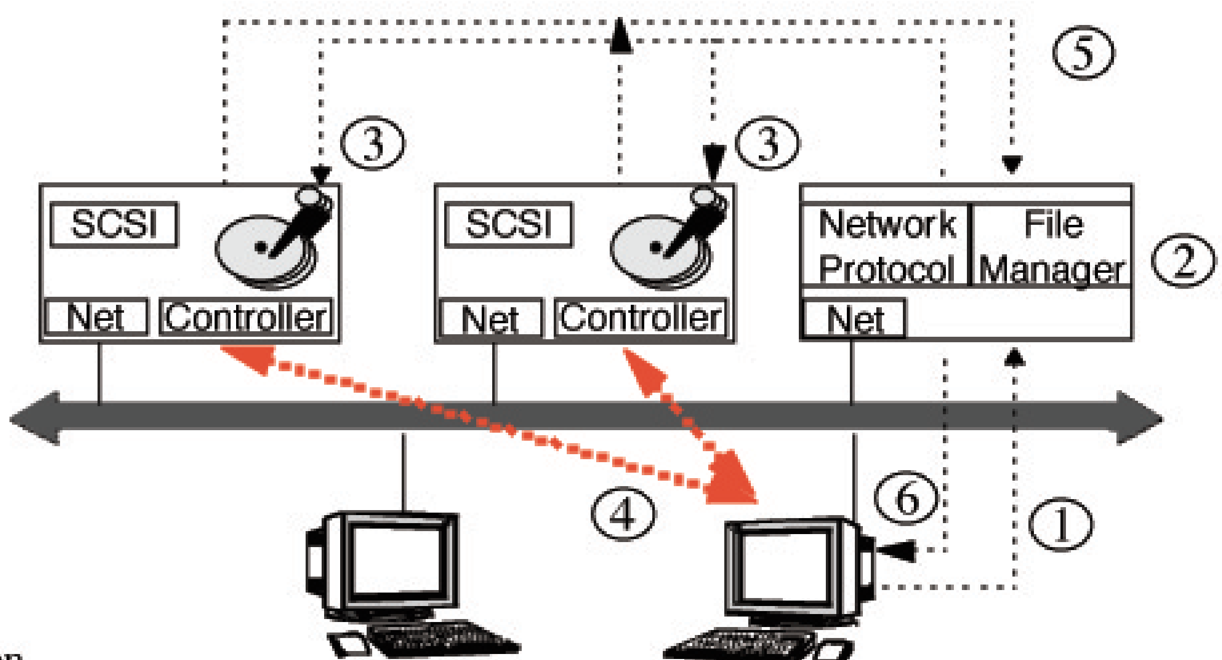
- **Symmetric multi-processor port of server to all client platforms**
  - Acquire locks (1), access metadata and data, release locks (3)
  - E.g. GFS/Sistina
- **But, bugs in heterogeneous clients & RW access to all storage**



panasas

Out-of-band 2: Metadata Server

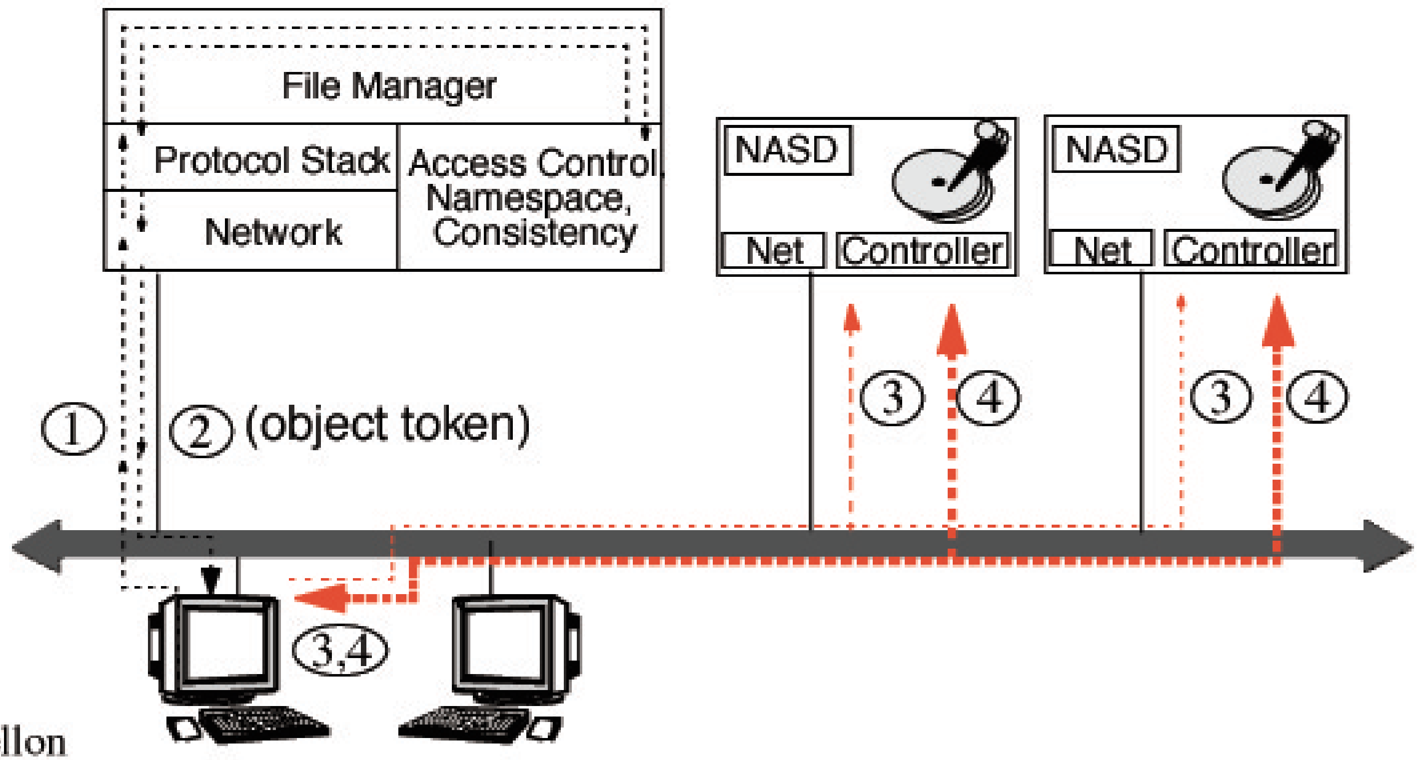
- **Central metadata server mediates access to storage**
  - Request (1) DMA (2) of data (3) to client (4) by server (5, 6) E.g. HPSS
  - Can cache readonly metadata at client, directly access storage E.g. High Road, SANergy, DirectNFS
- **But, metadata changes, including allocation, centralized**



panasas

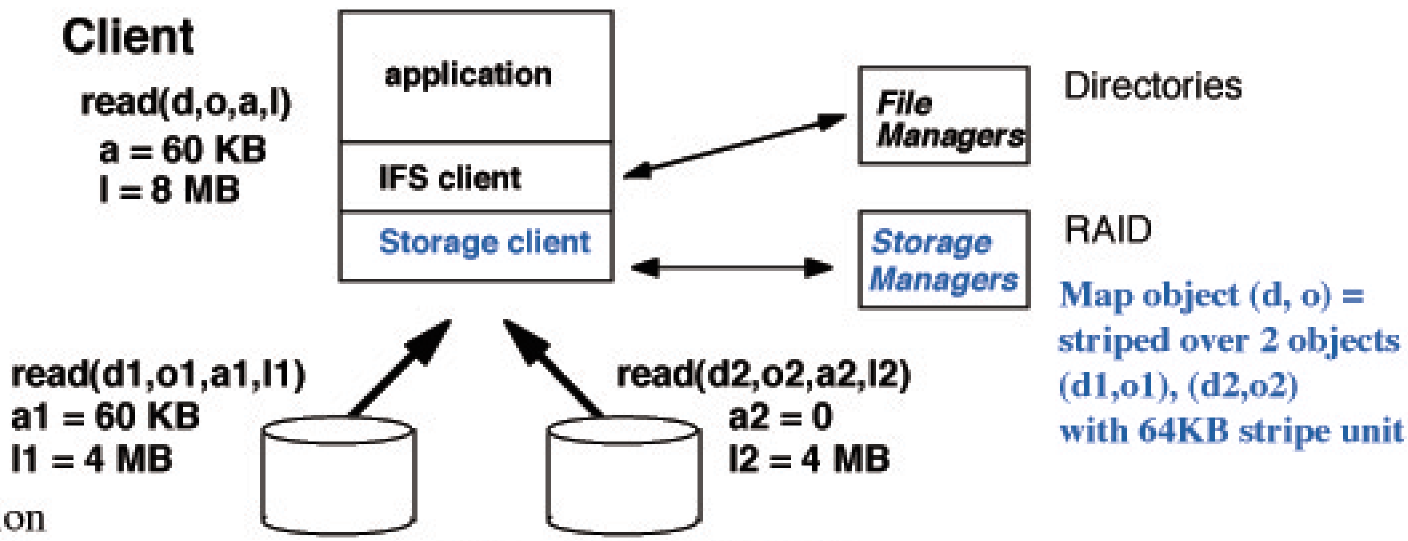
Out-of-band 3: Object Storage

- **File/object storage management in storage device (inode-like)**
  - Request (1, 2) and cache rights to read/write/extend objects on disks (3, 4)
    E.g. CMU NASD, Lustre
- **But, changes in storage device standards (integrated solutions)**



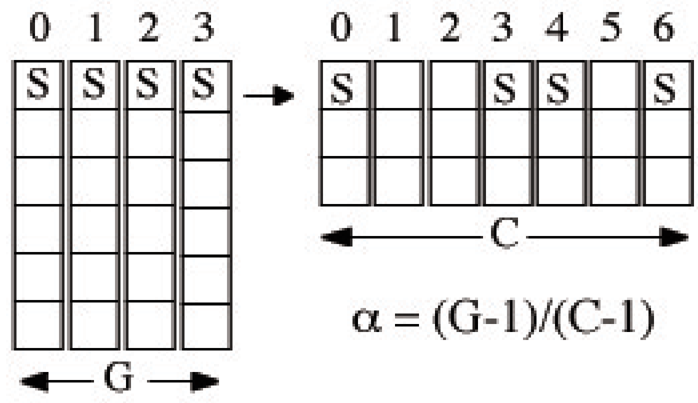
Incremental Growth

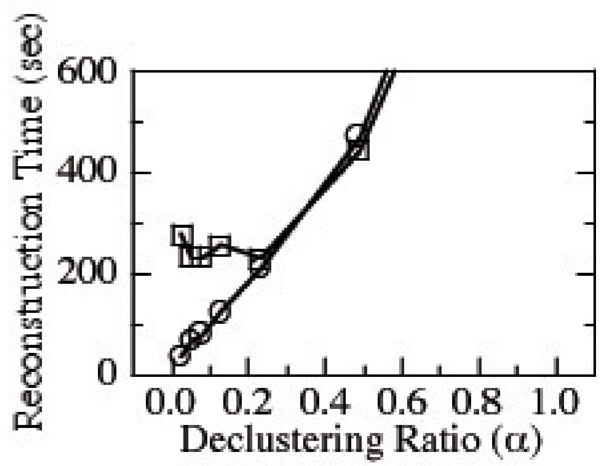
- **Computer science's duct tape: a level of indirection per object**
  - RAID remapping already there but hidden from file system and shared over many unrelated objects
- **Striping/RAID representation should be dynamic, file-specific**
  - Escrow capacity, defer allocation for fast path efficiency
  - Cache coherent, on-the-fly remapping for balancing and incremental growth
  - Embed representation in object attributes, extensible for QoStorage specification



Scaling Reliability

- **Flexible allocation enables declustering of redundancy groups**
- **Evenly distribute groups in larger arrays, reducing recovery work per disk**
- **Couple with integral spare space & *failure recovery time, R, is linear in declustering ratio (~G/C)***
- **MTTDL = K/(C*R), inversely proportional to size and recovery time becomes, *MTTDL = K/G, independent of size***
- **Requires XOR and net bandwidth to scale with number of devices, and declustered server failover**

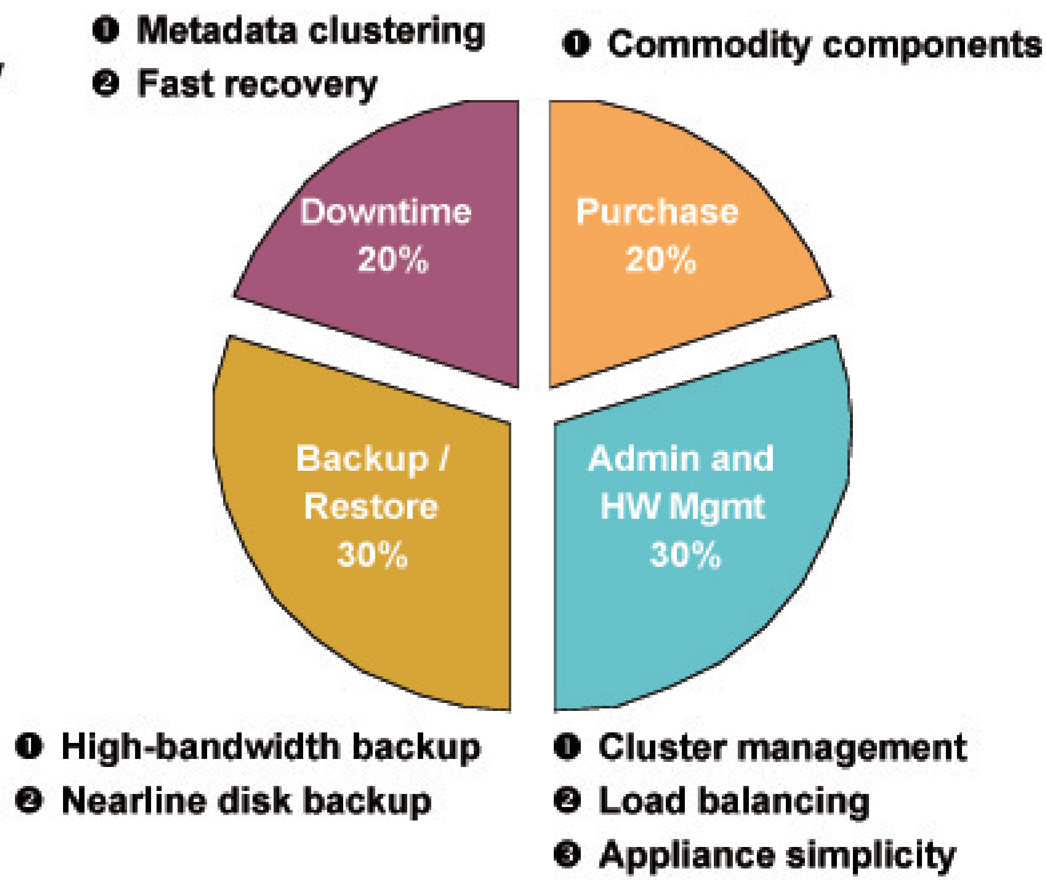




panasas

Total Cost of Ownership Control

- **Cost**:
  Commodity components lower capital costs of primary and nearline disk storage
- **Effort**:
  Appliance-like simplicity, cluster-wide management commands, automatic balancing, and per-file customization by performance API or policies
- **Recovery**:
  Exploit speed to reduce backup time, recovery time
  Transparent failover of metadata server



Source: Gartner Group, May 2001

panasas

Scaling Metadata Service

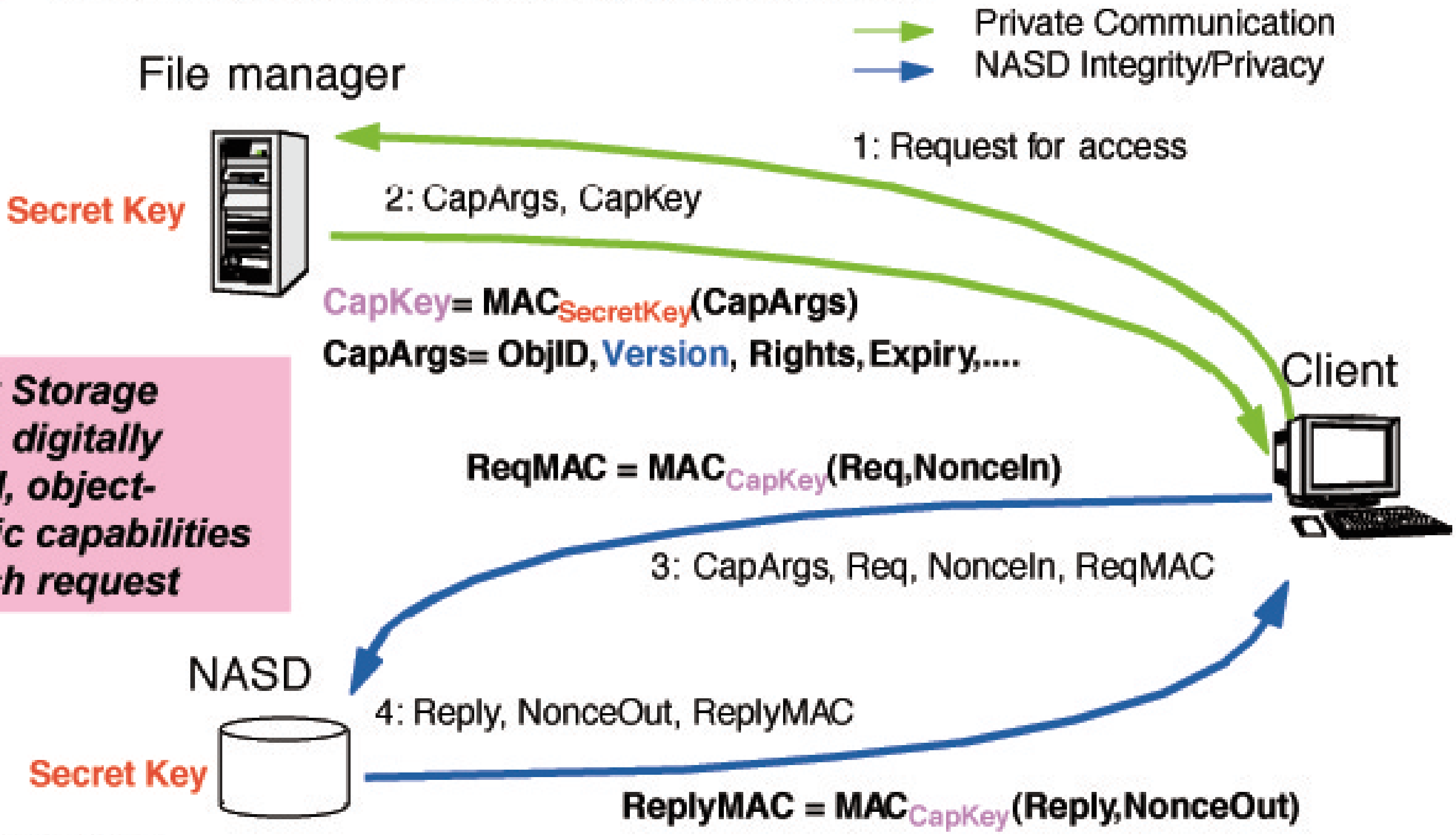
- Command processing of most operations in storage (or clients) offloads 90% of small file/productivity workload from servers

NFS Operation	Count in top 2% by work (K)	File Server (SAD)		DMA (NetSCSI)		Object (NASD)	
		Cycles (B)	% of SAD	Cycles (B)	% of SAD	Cycles (B)	% of SAD
Attr Read	792.7	26.4	11.8	26.4	11.8	0.0	0.0
Attr Write	10.0	0.6	0.3	0.6	0.3	0.6	0.3
Data Read	803.2	70.4	31.6	26.8	12.0	0.0	0.0
Data Write	228.4	43.2	19.4	7.6	3.4	0.0	0.0
Dir Read	1577.2	79.1	35.5	79.1	35.5	0.0	0.0
Dir RW	28.7	2.3	1.0	2.3	1.0	2.3	1.0
Delete Write	7.0	0.9	0.4	0.9	0.4	0.9	0.4
Open	95.2	0.0	0.0	0.0	0.0	12.2	5.5
Total	**3542.4**	**223.1**	**100**	**143.9**	**64.5**	**16.1**	**7.2**

panasas

Storage Access Security

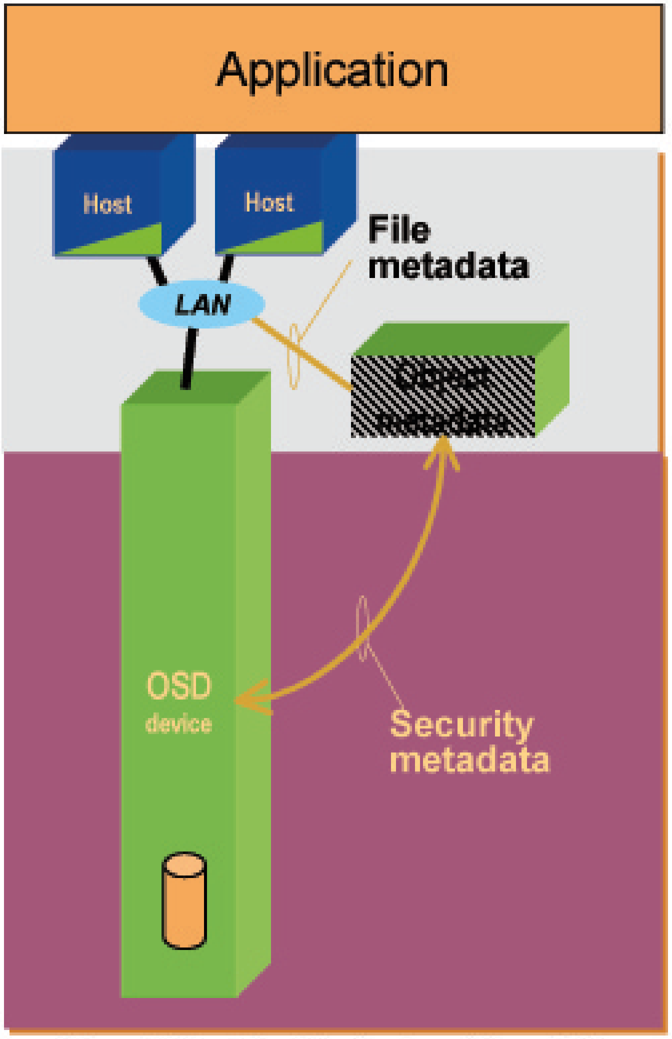
- **State of art is VPN of all out-of-band clients, all sharable (meta)data**
  - Accident prone and vulnerable to subverted client



Object Storage uses a digitally signed, object-specific capabilities on each request

Object storage: When competitive?

- **NSIC first draft collaboration (96-99)**
  - CMU, IBM, Seagate, HP, STK, Quantum
  - CMU releases software (Extreme NASD)
- **SNIA OSD TWG and ANSI T10 OSD WG**
  - Featured in SNIA technical council shared storage model taxonomy
  - SNIA TWG chaired by Intel and Ciprico
    T10 WG chaired by Ralph Weber
  - Sun, STK, Seagate, MTI, HP, Panasas
  - Strong leveraging of iSCSI draft standard
- **Rumors of possible commercial variations**
  - IBM Storagetank (FAST), EMC (Eweek), Sun (Byte&Switch)
  - Stay tuned …



Storage Networking Industry Association, 2001

Next Generation Agile Storage

- Out-of-Band for bandwidth, parallelism, fast recovery
- Commodity components integrated for cost effectiveness
- Shared-nothing clustering scales
- Self-describing data for dynamic, file-specific representation
- Object Interfaces encourage CPU, memory & link speed to scale in proportion to spindles

panasas

www.panasas.com